

データグリッド技術

松田秀雄、大川剛直 (阪大 情報科学研究科)

安永照雄 (阪大 遺伝情報実験センター)

古舘丈裕、若月謙太郎 (日立ソフトウェアエンジニアリング)

北島正人 (富士通九州システムエンジニアリング)

河村元 (アズテックシステム)

磯山正治 (蛋白質研究奨励会)

藤博幸 (京大化研)、佐藤仁則、石川元一 (三菱スペース・ソフトウェア)

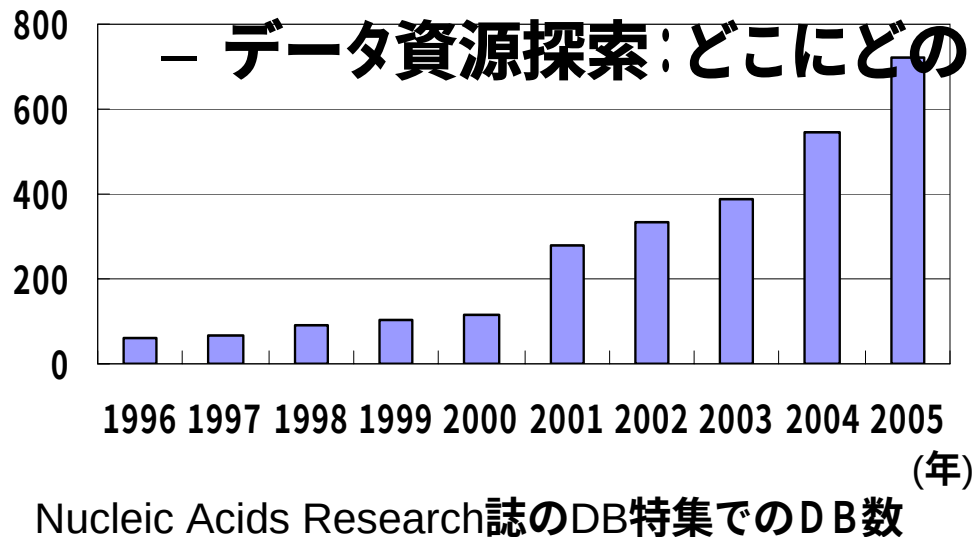
齋藤裕介、国府田明弘 (日本ヒューレットパッカー)

伊達進 (阪大 情報科学研究科)

下條真司 (阪大 サイバーメディアセンター)

- 広域に分散した大規模データへの対応の必要性：
 - 高エネルギー物理学、天文学、気象、ゲノム、プロテオームなど
- データサービスの多様化：キーワード検索以外の検索の必要性

- 構造類似性検索：データの持つ構造に対する知識が必要
- 関連性検索：異分野の概念同士の対応関係が必要
- データ資源探索：どこにどのようなデータ資源が存在するか



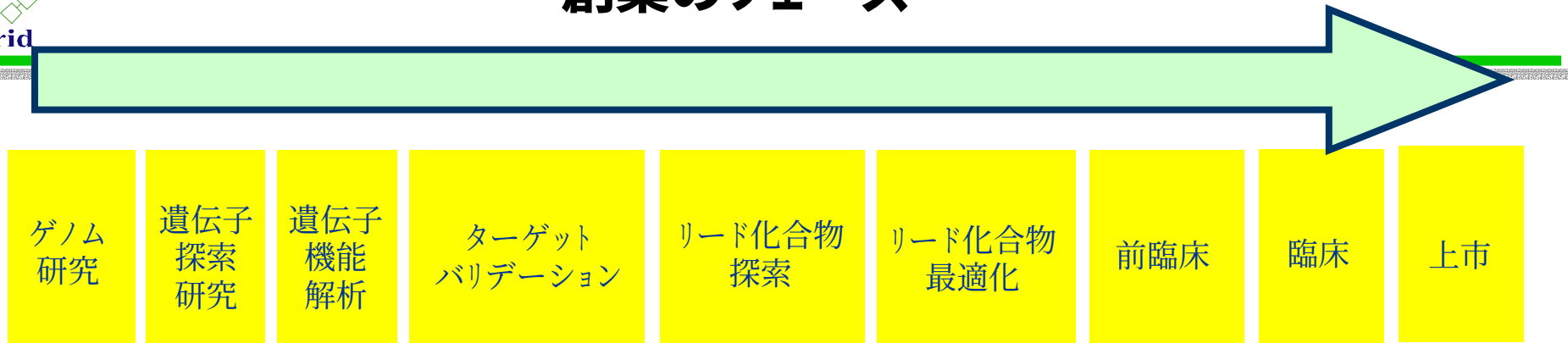
データの種別	DBの数
DNA	87
RNA	29
タンパク質	94
ゲノム	58
マッピング	29
タンパク質立体構造	18
文献情報	43
その他	153

合計：511

Bio DB Catalog (DBCAT)への登録数：
<http://www.infobiogen.fr/services/dbcat/>

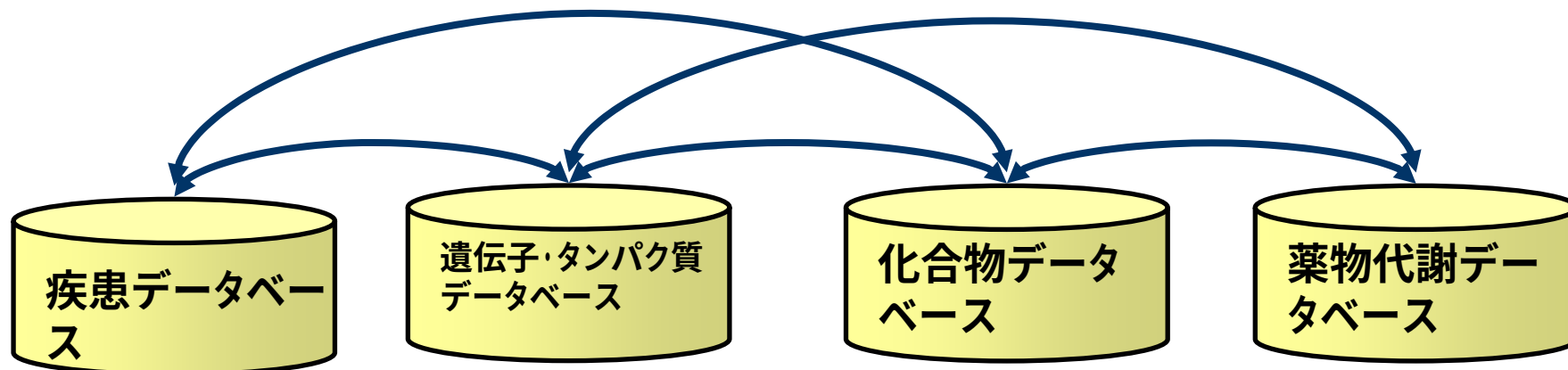
- 物理的に1箇所にデータをまとめて単一のデータベースとして運用
 - 検索効率是最も高いが、データ更新の問題があり現実的でない
(GenBankは毎日平均30MB程度の更新データ。年平均2回のスキーマ更新)
- インターネット上で仮想的に統合
 - リンク統合 (例SRS):
 - (長所) 特別なツールなしに容易に利用できる。
 - (欠点) 基本的にキーワード検索のみしかなく意味を表現できない。
 - ウェブサービスによる統合 (例DAS):
 - (長所) あらゆる検索方法をサービスとして結び付けられる
 - (欠点) サーバごとの情報を統一する規則がない (異義語の問題など)

創薬のフェーズ



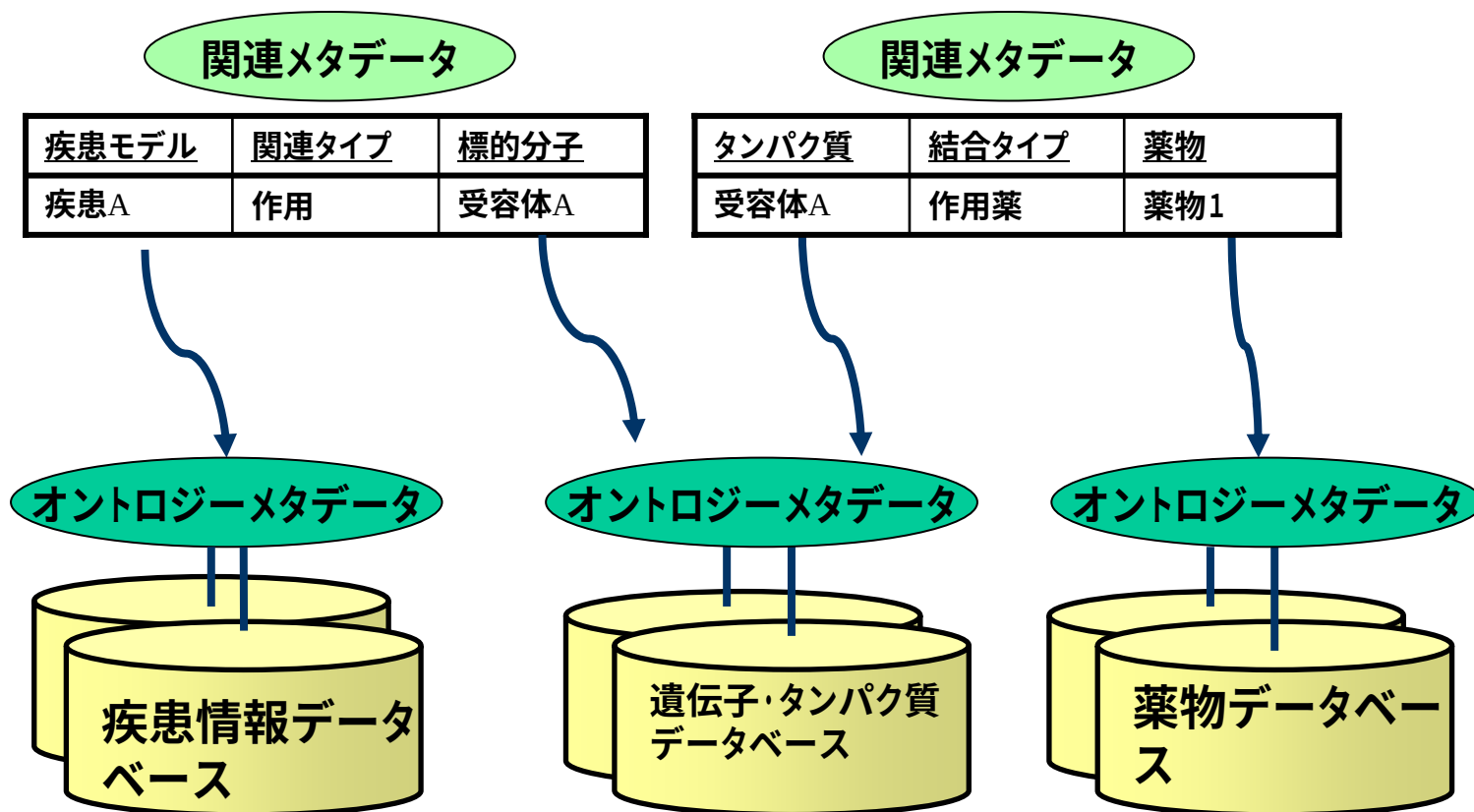
アプリケーションへの負担

多様な関連性→膨大な組合せ
用語の違い、表記のゆれ

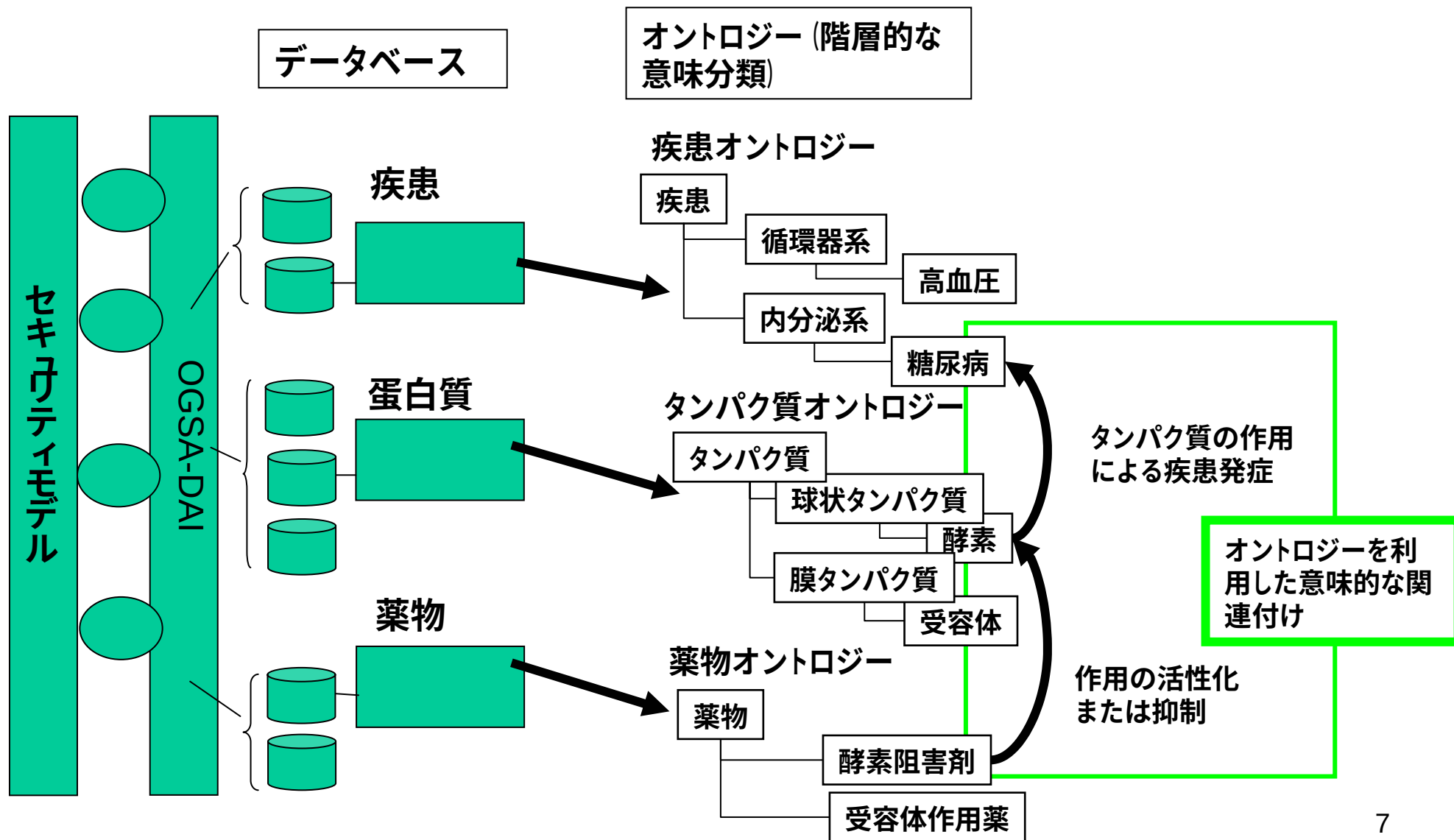


- **2階層のメタデータを介してデータベースを参照する。**
 - **オントロジーメタデータ**: オントロジーにより概念体系を明確化し、各分野の実体 (例: タンパク質、薬物など) ごとの用語の違いやあいまい性を吸収
 - **関連メタデータ**: 実体間の意味的な関連を記述
- **メタデータ導入の利点**
 - 複数のデータベースで、データ間の関連全てを直接記述すると、データ数の組合せに比例する膨大な数になる (**組合せ爆発**)。
 - メタデータを介することにより、データ間の対応を階層的な分類間の関連にまとめることができる。

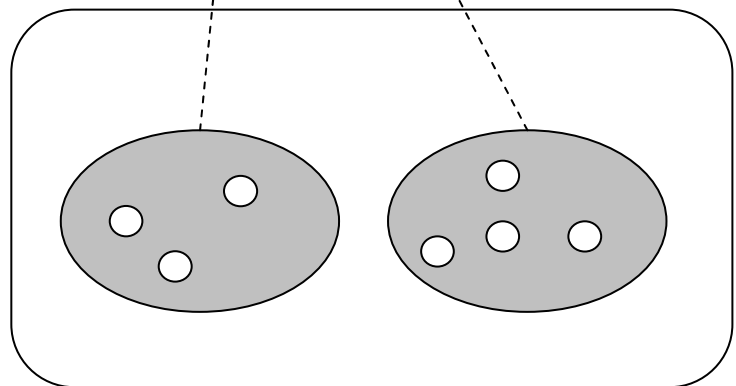
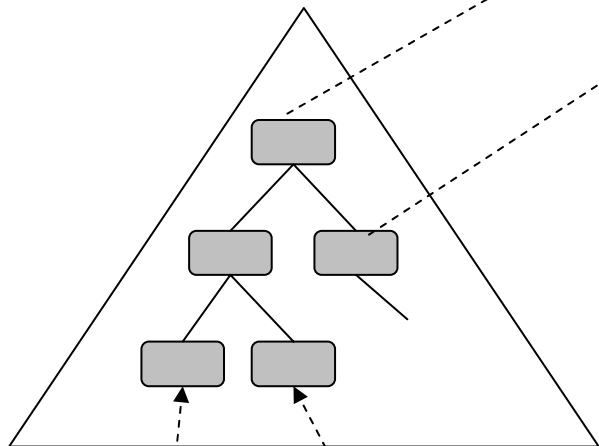
メタデータによる異分野データベースからの情報統合



オントロジーによる異分野データベース間の関連付け



タンパク質機能 オントロジー

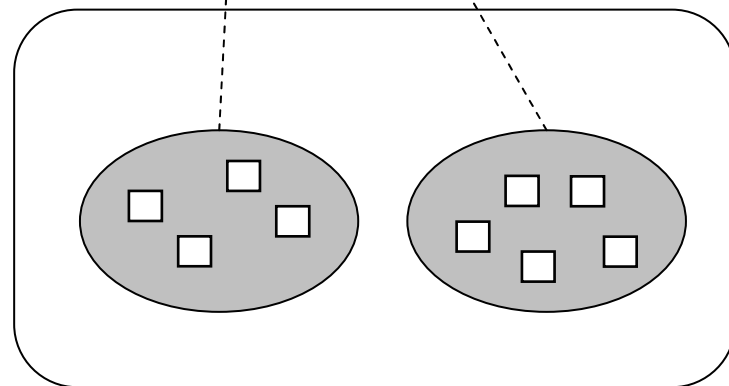
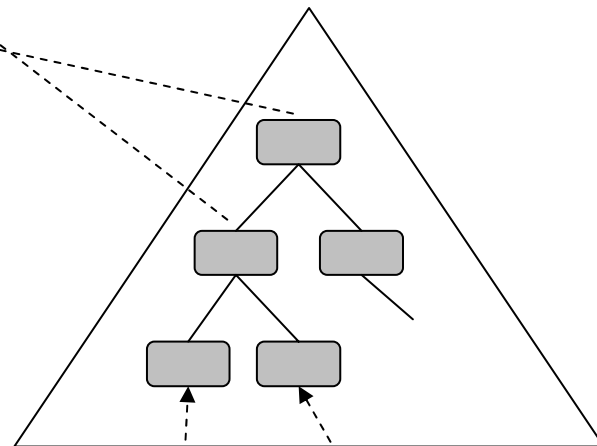


タンパク質

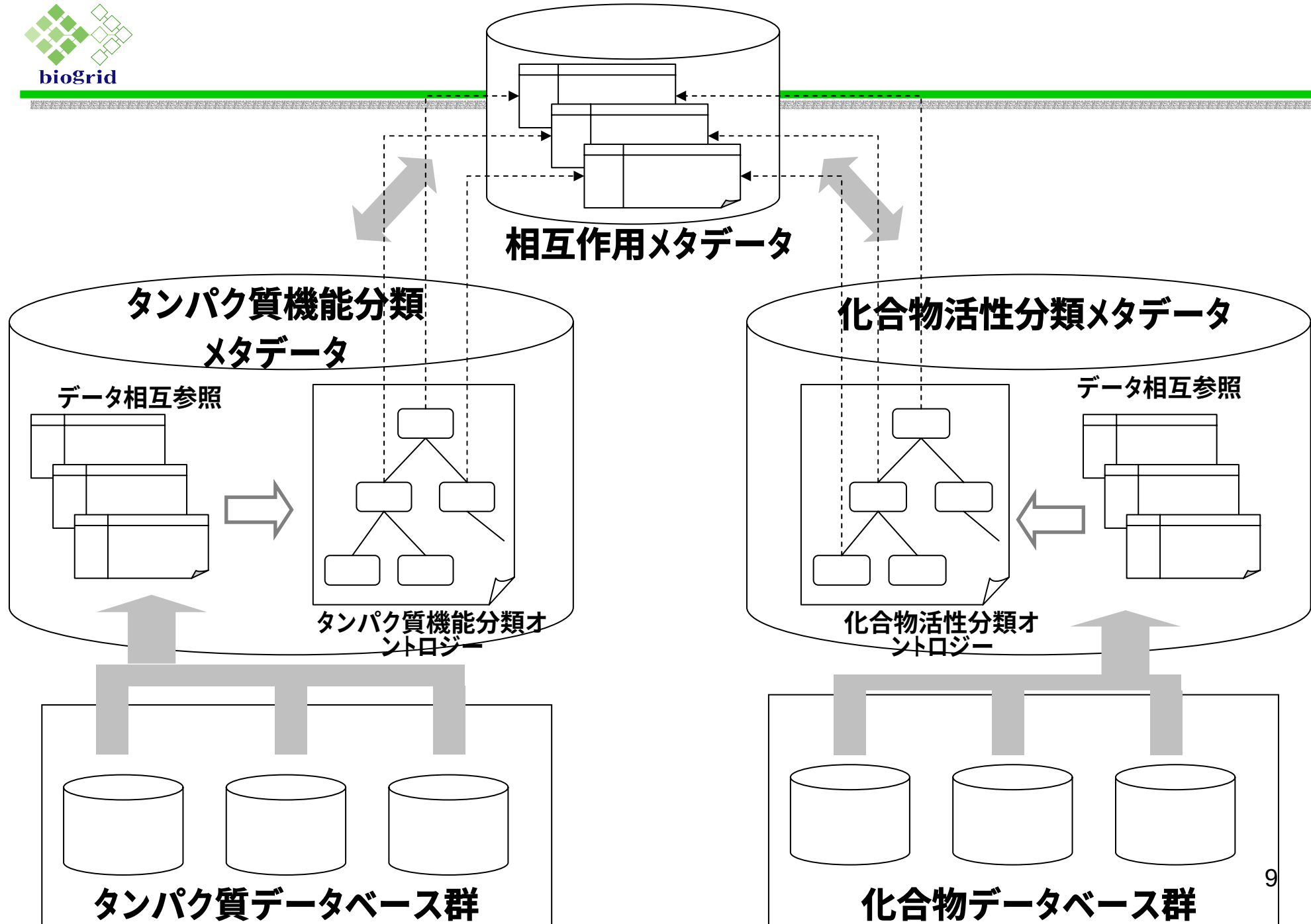
作用薬
拮抗薬
阻害薬

タンパク質・化合物
相互作用

化合物活性 オントロジー



化合物



疾患

Medical Encyclopedia

3K entries

PRF DB

(蛋白質研究奨励会)

パスウェイ

(生体分子間
ネットワーク)

KEGG

18K entries

タンパク質

SwissProt 163K entries,
60M amino acids

PIR 283K entries,
96M amino acids

PDB 28K protein structures

ゲノム

Ensembl, DDBJ

Human 24K genes

3.3G nucleotides

化合物

MDDR 150K compounds

PubChem, Ligand Info

1.8M compounds

相互作用

ENZYME

GPCR-DB

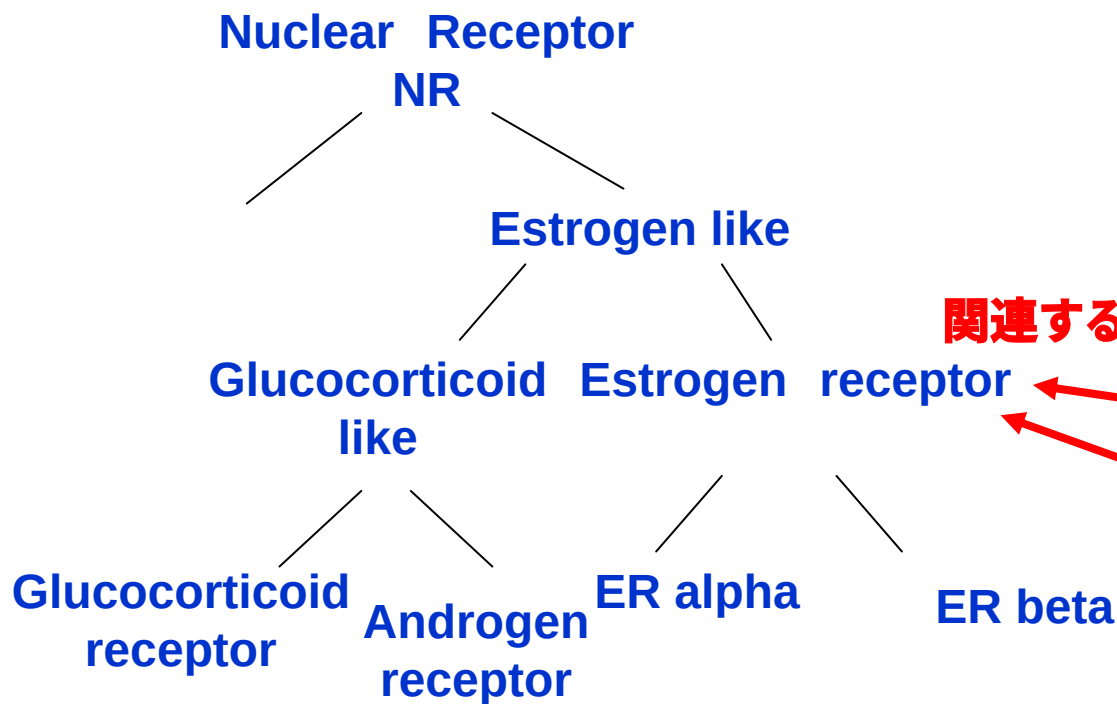
NucleaRDB

LGIC DB

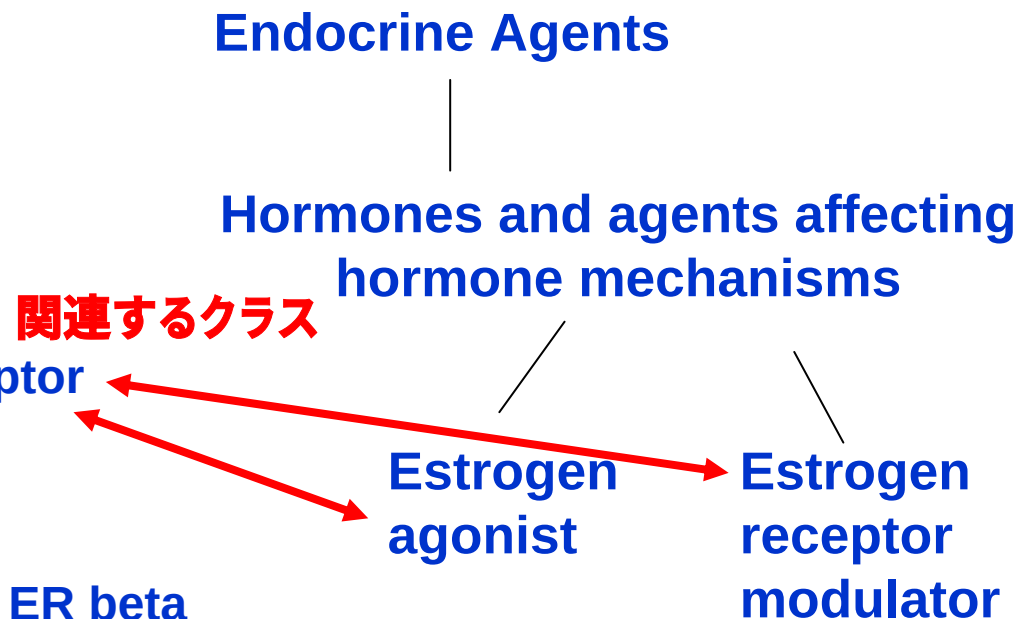
- Ligand.Info (<http://ligand.info/>) : **以下の公共化合物データベースから抽出**
 - ChemBank subset 2,344件
 - ChemPDB subset 4,009件
 - KEGG subset 10,005件
 - Anti-HIV NCE subset 42,689件
 - Drug-likeness NCI subset 192,323件
 - Not annotated NCI subset 15,237件
 - AKos GmbH subset 544,391件
 - Ligand.InfoAsinex Ltd. subset 348,276件
 - Tim Tec subset 7,500件

計1,159,274化合物
- PubChem (<http://pubchem.ncbi.nlm.nih.gov/>)
 - 711,361**件の化合物構造情報**
 - High Thorough Screening **による344件のBio Assay系のデータ**
- Drug Bank(<http://redpoll.pharmacy.ualberta.ca/drugbank>)
 - 256**件の化合物に関する薬物商標名**

タンパク質 の機能分類 (Gene Ontology)



化合物 の活性分類 (MDL Drug Data Reportの活性情報から抽出)



- タンパク質・化合物を分類したオントロジーで、対応可能な概念クラスを階層構造をたどりながら探索し、クラス間の関連を抽出する。
- 実際の分子間相互作用データで検証していく

相互作用メタデータの生成 (配列情報解析)

1. BLASTによって相同性を判定

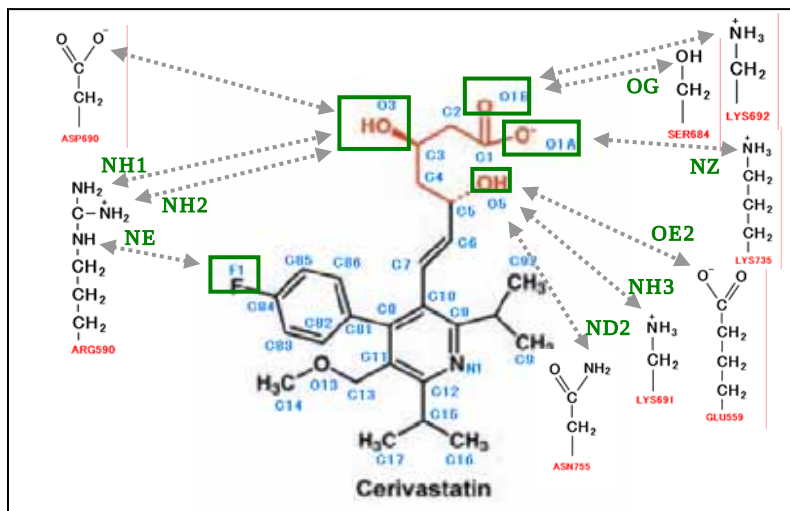
Query: 1 MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEVILLHVIDEREIKKRDIFSLLLGVA 60
M M++K+L+PTDFSE A A++ + ++ EVILLHVIDE +++ L+ G +

Subjct: 1 MIFMFRKVLFP TDFSEGAYRAVEVF EKRNKMEVGEVILLHVIDEGTLEE-----LMDGYS 55

Query: 61 GLNKSVEEFENELKNKLTEEA~~KN~~KMENIKKELEDV--GFKVKDII~~V~~VGIPHEEIVKIAED **118**
 + E ++K KL EEA K++ +E++ V+ II GIP +EIVK+AE+

Sbjct: 56 FFYDNAEIELKDIKEKLKEEASRKLOEKAEVVKRAFRANKVVRTIIRFGIPWDEIVKVAEE 115

2.化合物との相互作用部位を残基モチーフに適用



類似活性部位を持つモチーフの抽出

MYADLILEAE LAVEEQPDQGV EGPLGTPG
SYAELMPVERILPVEOPLAVEPKLETPV

同様の薬剤活性を示す残基配列は保存されている

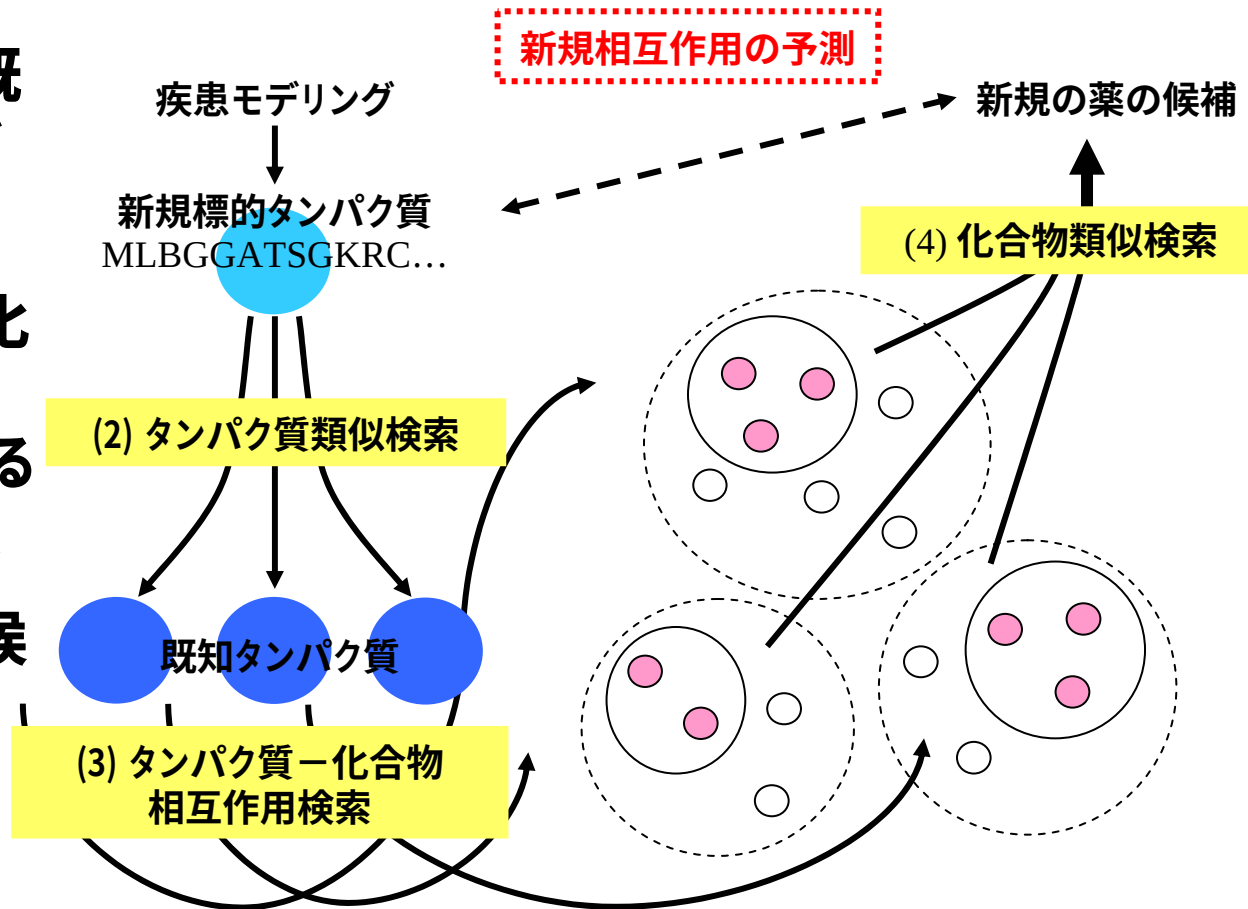
→活性情報のメタデータ化

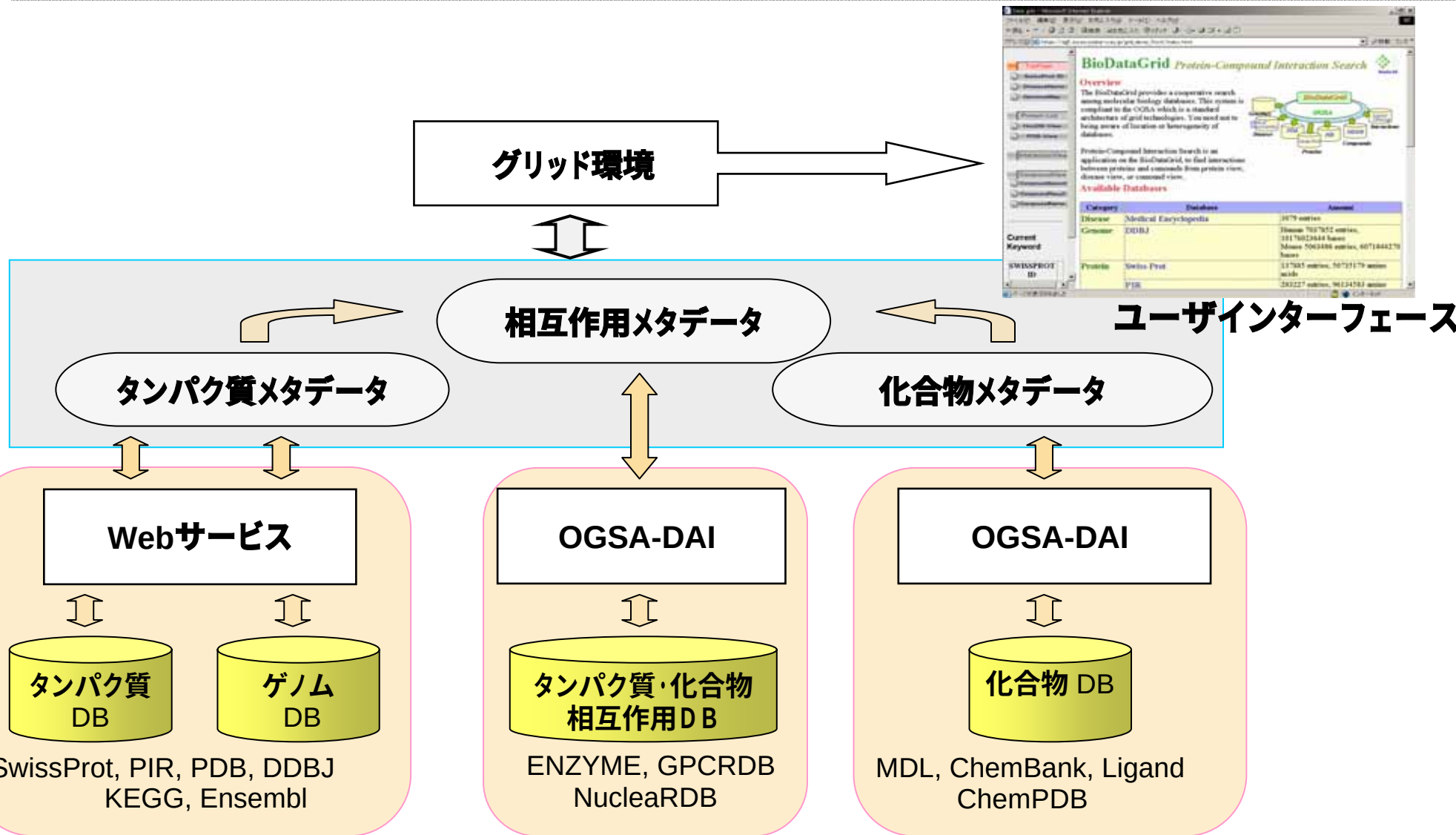
1. 疾患に関連する新規の標的タンパク質を検索する。

2. 標的タンパク質に類似した既知タンパク質を、BLASTなどを用いて検索する。

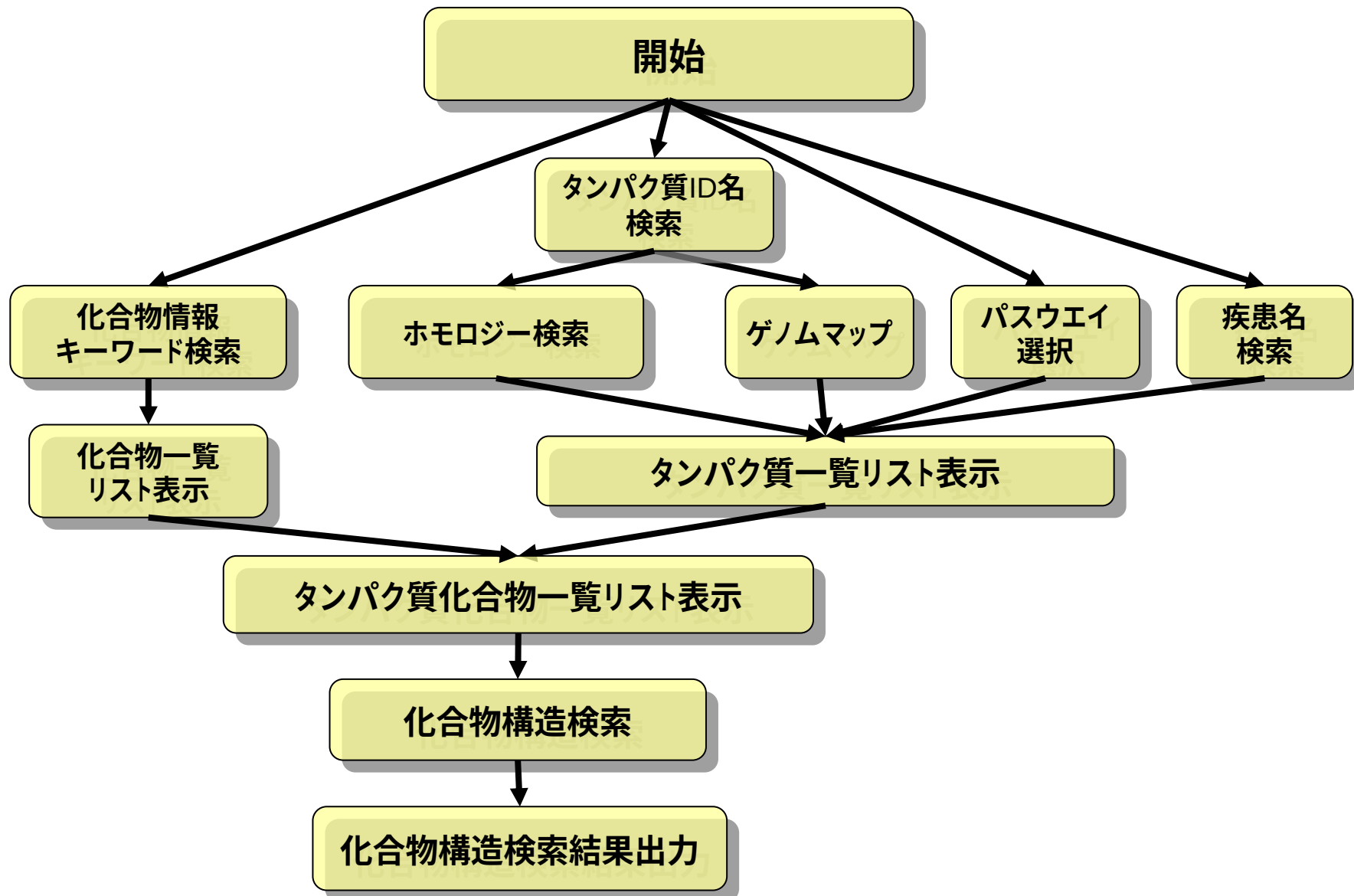
3. 既知タンパク質に結合する化合物を、タンパク質と化合物の相互作用データから求める

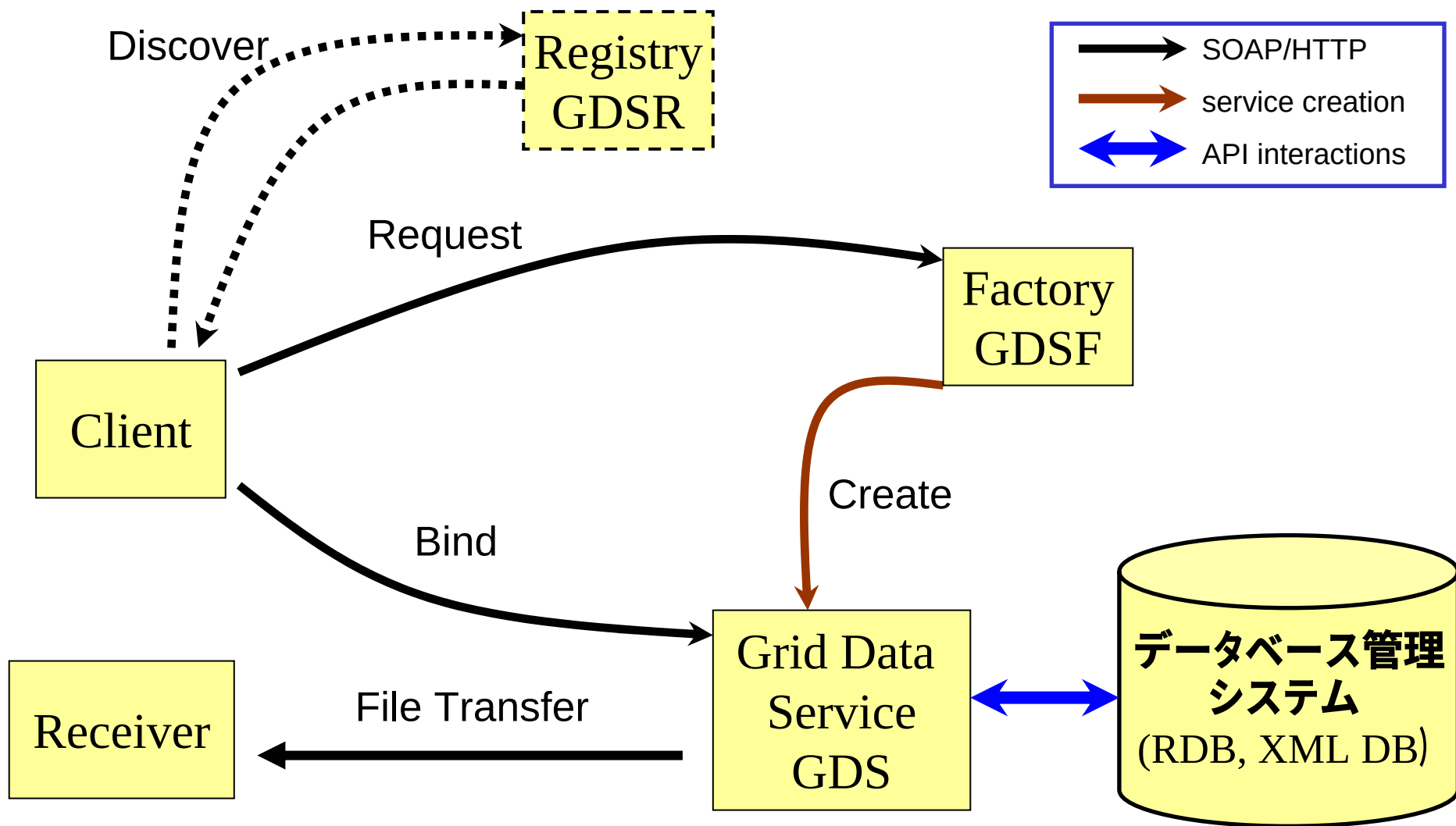
4. 化合物類似検索を用いて検索された化合物から、薬の候補となる化合物を検索する。

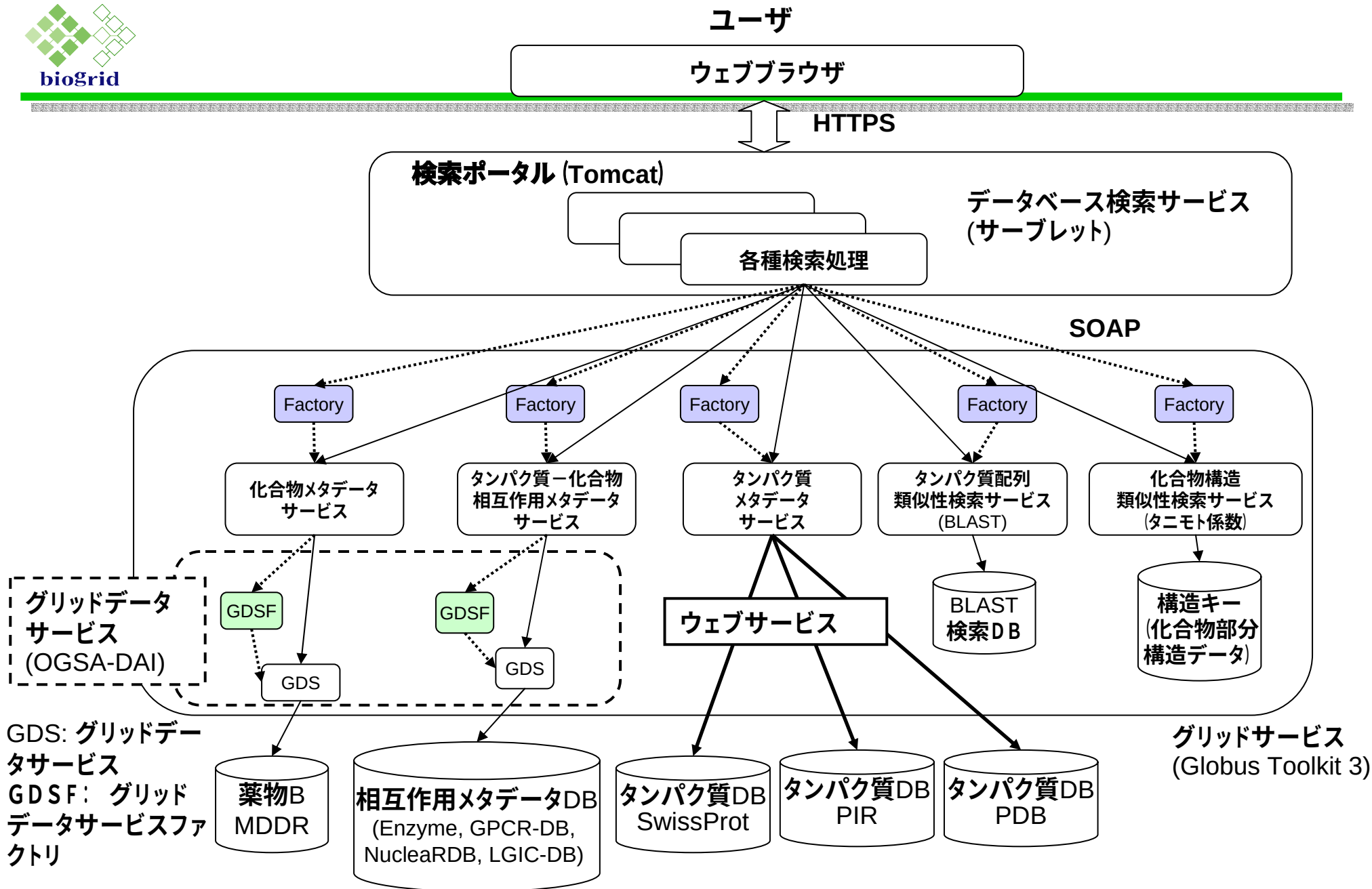




現状のデータグリッドシステムの状態遷移図



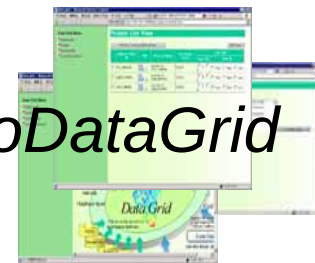


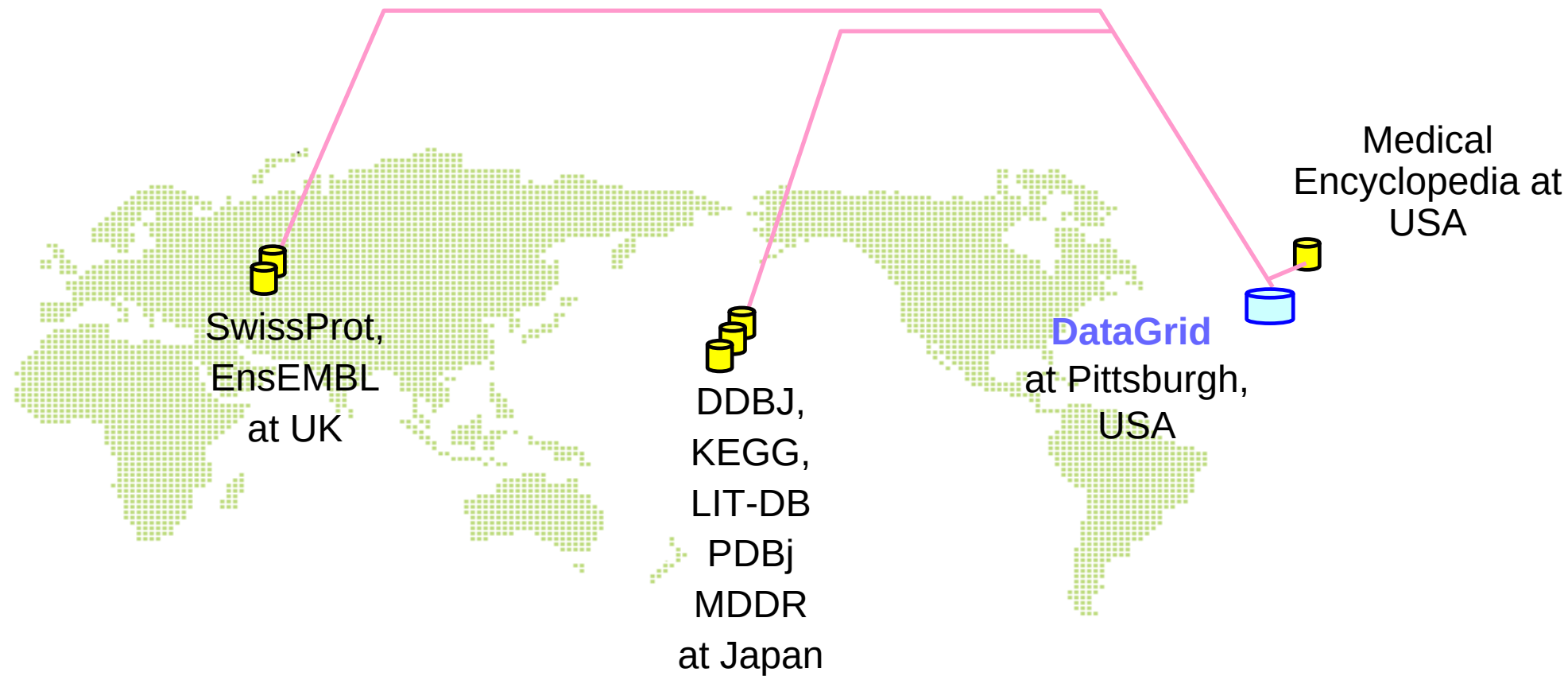




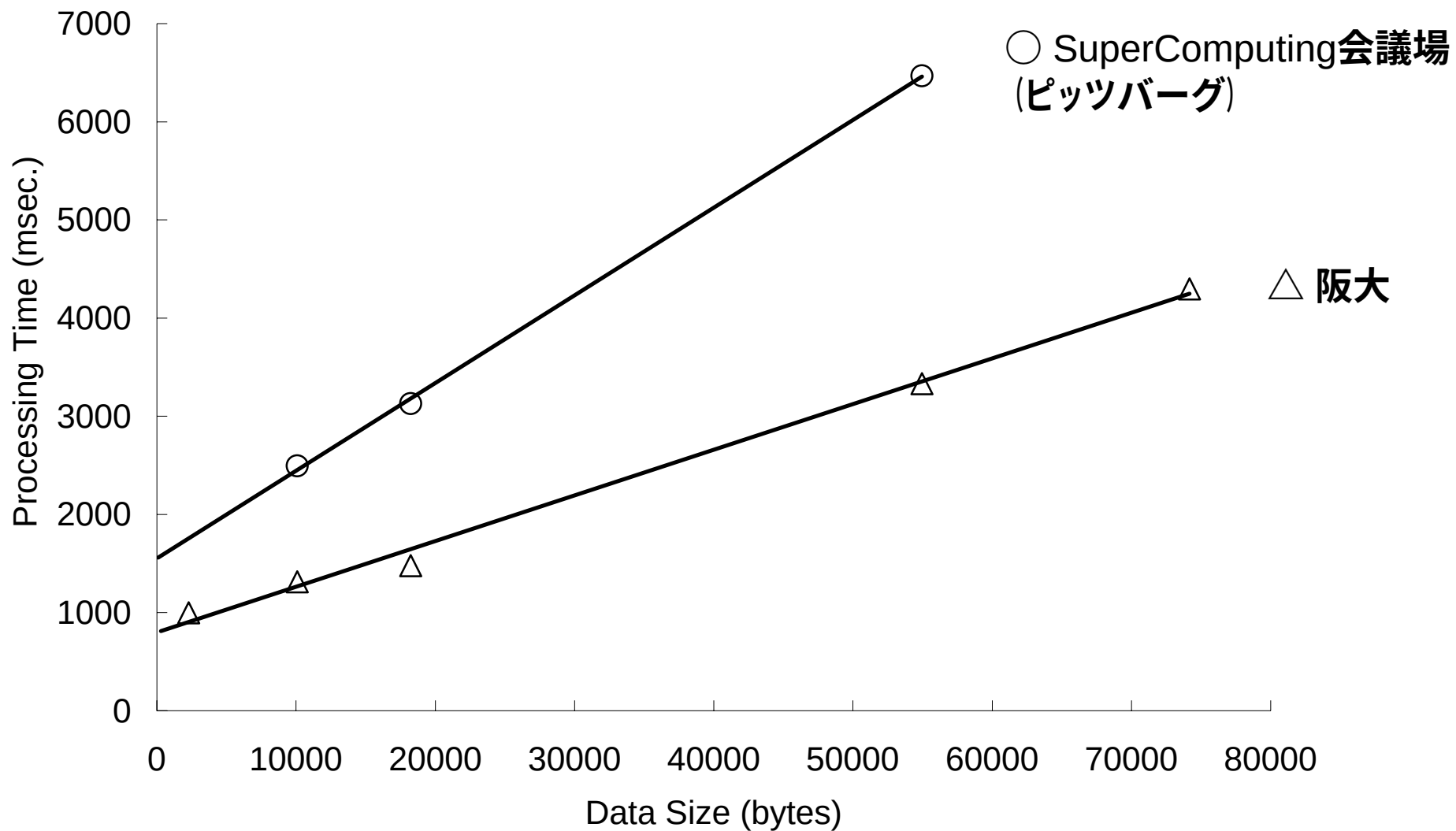
Category	Database	Amount
Disease	Medical Encyclopedia	3079 entries
Genome	Ensembl	Human 24,195 genes, 3,272,187,692 bases
	DDBJ	Mouse 28,055 genes, 2,932,352,227 bases
Protein	Swiss-Prot	163,235 entries, 59,631,787 amino acids
	PIR	283,416 entries, 96,216,763 amino acids
	PDB	27,969 structures
Compound	MDL Drug Data Report (MDDR-3D) Ver. 2004.1	147969 entries
Interaction	Ligand Ontology	ENZYME, GPCR-DB, NucleaRDB, LGIC-DB

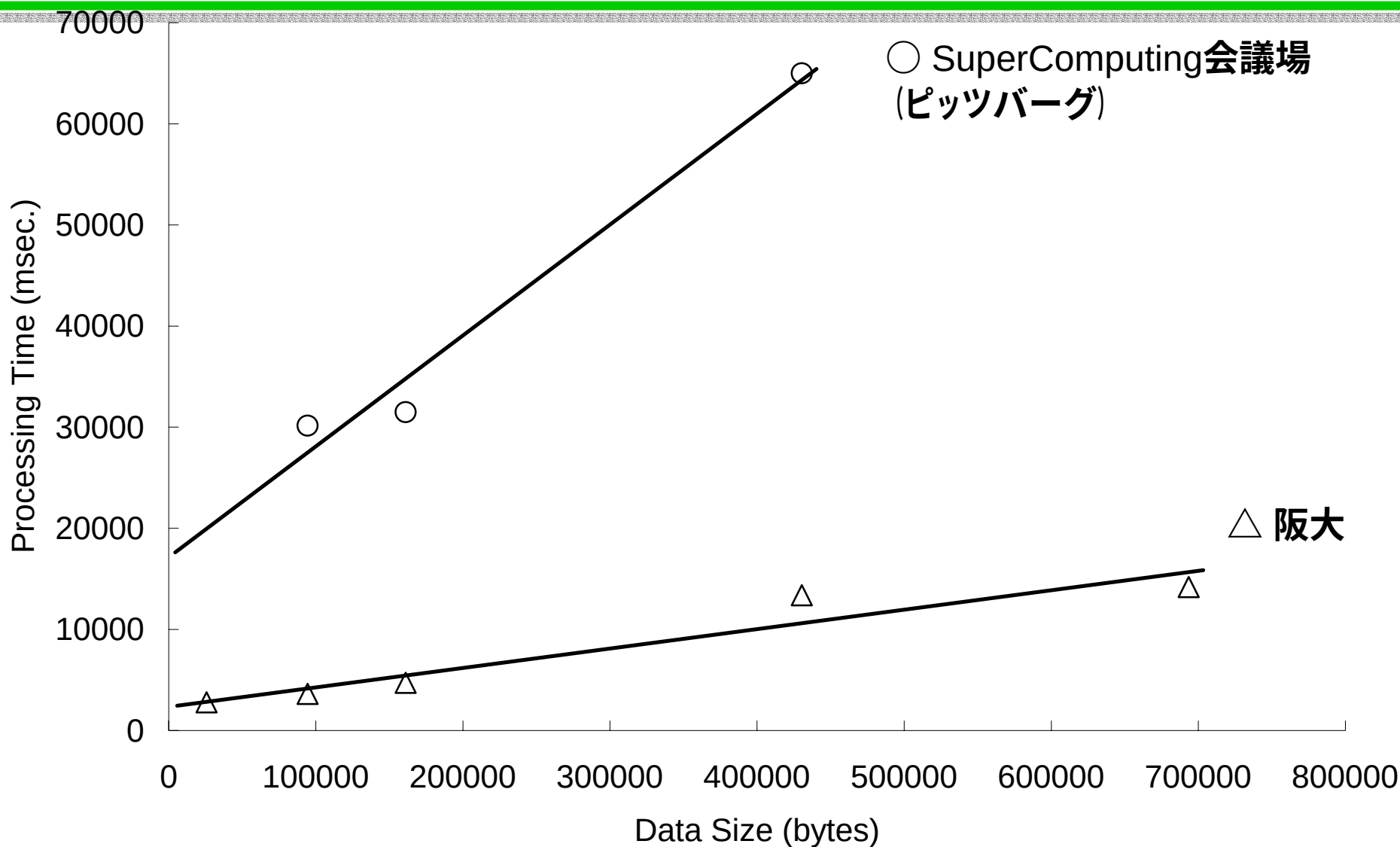
BioDataGrid





地理的に分散したウェブサービスをデータグリッドで合成





- **OGSA-DAIを使った、メタデータによる情報統合システムのプロトタイプを開発**

SC2004でウェブサービス版 (データベースのサービス元へアクセス) のデモを行った。

- **現在、情報統合の対象にしているのは16個のDB**

PRF LITDB, Medical Encyclopedia, KEGG, DDBJ, EnsEMBL, SwissProt, PIR, PDB, Enzyme, GPCR-DB, NucleaRDB, LGIC-DB, MDDR, ChemBank, ChemPDB, NCI

- **課題**

- **ワークフローの実現**
- **組合せ爆発の抑制**